Text Mining – Homework 2

Executive Summary:

The main goal of this project was to classify textual information into authenticity and sentiment (positive or negative) categories. This included news articles and restaurant reviews. Lemmatization, text normalisation, tokenization, stopword removal, and other crucial data preprocessing steps were performed at the beginning of the procedure. The textual information was then converted into numerical features that enabled machine learning through a variety of textual techniques, including binary encoding, term frequency counts, unigrams, bigrams, and term frequency-inverse document frequency (TF-IDF) representations.

With the help of three text vectorization techniques—Unigram Term Frequency Vectorization, Unigram and Bigram Term Frequency Vectorization, and Unigram TF-IDF Vectorization—a total of 12 machine learning models were trained and tested for this project. Both sentiment analysis and authenticity classification of these models were assessed.

Multinomial Naive Bayes (MNB) and Support Vector Machine (SVM) stood out among the models. It should be noted that the SVM model performed admirably, earning an accuracy score of 0.964 for sentiment analysis using Unigram TF-IDF Vectorization. Being cautious is important because a model that performs well on training data but may not generalise well to new, untested data may be overfitting. To evaluate its generalisation potential, more validation, cross-validation, and hyperparameter tuning are required.

The MNB model with Unigram TF-IDF Vectorization performed best for authenticity classification, achieving an accuracy score of 0.79. The ability of this model to distinguish between real and fake data was strong.

A feature analysis was carried out to find key words for sentiment and authenticity classification in addition to model performance evaluation. Error analysis highlighted instances where the models incorrectly classified data, revealing areas for potential model improvement.

This project presents a structured method for analysing textual data that includes text vectorization, data preprocessing, and the creation and assessment of machine learning models. Particularly in applications like fake news detection and sentiment analysis of restaurant reviews, the ability of these models to categorise sentiment and authenticity has significant value. Incorporating feature analysis, error analysis, and evaluation metrics offers crucial insights into the performance of the model and possible areas for improvement.

a. The models you have built and tested, and which one is the best for each task

The SVM model with Unigram TF-IDF Vectorization displayed exceptional performance for sentiment analysis with an accuracy score of 0.964.

With an accuracy score of 0.79, the MNB model with Unigram TF-IDF Vectorization outperformed the competition in the authenticity classification.

b. Top 10 features for each category in each task, for each algorithm. Explain if these features make sense to you, and what they mean regarding sentiment and authenticity.

Feature Analysis for Sentiment data

Model	Vectorisation Method	Top 5 Features	Worst 5 Features
MNB	Unigram Term Freq Vectorizer	["amazing", "best", "love", "friendly", "great"]	["bland", "came", "drink", "minute", "said"]
MNB	Unigram and Bigram Term Freq Vectorizer	["amazing", "best", "love", "friendly", "great"]	["bland", "came", "drink", "minute", "worst"]
MNB	Unigram TF-IDF Vectorizer	["amazing", "best", "great", "love", "fresh"]	["came", "drink", "said", "worst", "bad"]
SVM	Unigram Term Freq Vectorizer	['best', 'need', 'great', 'price', 'little']	['table', 'dish', 'drink', 'said', 'asked']
SVM	Unigram & Bigram Term Freq Vectorizer	['best', 'delicious', 'great', 'need', 'fresh']	['asked', 'dish', 'said', 'cold', 'place']
SVM	Unigram TF-IDF Vectorizer	['best', 'great', 'amazing', 'friendly', 'need']	['place', 'dish', 'bad', 'asked', 'terrible']

Feature Analysis for Authenticity data

Model	Vectorisation Method	Top 5 Negative Features	Top 5 Positive Features
MNB	Unigram Term Frequency Vectorizer	["definitely", "recommend", "kind", "bring", "minute"]	["environment", "people", "pretty", "make", "bar"]
MNB	Unigram & Bigram Term Frequency Vectorizer	["kind", "bring", "definitely", "minute", "worth"]	["environment", "people", "cheese", "pretty", "make"]
MNB	Unigram TF-IDF Vectorizer	["minute", "recommend", "worth", "definitely", "bring"]	["people", "environment", "bar", "cheese", "home"]
SVM	Unigram Term Frequency Vectorizer	["went", "noodle", "great", "quality", "kind"]	["time", "make", "people", "indian", "gone"]
SVM	Unigram & Bigram Term Frequency Vectorizer	["went", "great", "best", "cold", "service"]	["best restaurant", "time", "indian", "make", "restaurant ever"]
SVM	Unigram TF-IDF Vectorizer	["cold", "ice", "plate", "minute", "waiter"]	["time", "people", "gone", "restaurant", "environment"]

In the context of sentiment and authenticity classification, the features discovered through analysis make sense. Positive words like "amazing," "best," and "love" indicate a positive sentiment in sentiment analysis, whereas words like "bland" and "worst" indicate a negative sentiment. Terms like "definitely," "recommend," and "kind" suggest genuine reviews, while

unusual or exaggerated terms may indicate fake content, according to criteria for authenticity. These features allow us to understand how particular words affect the classification of reviews without disclosing the actual feature terms, and they give us valuable insights into the textual cues that machine learning models rely on to distinguish sentiment and authenticity.

c. Each model's performance in accuracy, precision, recall, and F-measure

Model – Sentiment data

Model	Accuracy	Precision	Recall	F1 Measure
MNB Unigram TF Vector	0.857	0.79	0.92	0.85
MNB Unigram & Bigram	0.857	0.79	0.92	0.85
MNB Unigram TF-IDF	0.857	0.79	0.92	0.85
SVM Unigram TF	0.75	0.69	0.75	0.72
SVM Unigram & Bigram	0.75	0.80	0.75	0.77
SVM Unigram TF-IDF	0.964	0.92	1.0	0.96

Model – Authentic data

Model	Accuracy	Precision	Recall	F1-Score
MNB Unigram TF Vector	0.75	0.75	0.69	0.72
MNB Unigram & Bigram TF Vector	0.68	0.64	0.69	0.67
MNB Unigram TF-IDF Vector	0.79	0.82	0.69	0.75
SVM Unigram TF Vector	0.57	0.54	0.72	0.62
SVM Unigram & Bigram TF Vector	0.49	0.48	0.72	0.58
SVM Unigram TF-IDF Vector	0.49	0.48	0.72	0.58

d. Each model's error analysis to identify areas for improvement

Data:	Model Name	Incorrect Predictions
Sentiment	MNB: Unigram TF Vector	7
Sentiment	MNB: Unigram & Bigram	4
Sentiment	MNB: Unigram TF-IDF	4
Sentiment	SVM: Unigram TF	7
Sentiment	SVM: Unigram & Bigram	4
Sentiment	SVM: Unigram TF-IDF	1

SVM with Unigram TF-IDF Vectorization stands out as the most accurate sentiment model in the analysis, with only one incorrect prediction. Regardless of the vectorization method used, MNB models have slightly higher error rates, with 4 to 7 incorrect predictions on average. With 4 incorrect predictions each, SVM: Unigram & Bigram and MNB: Unigram & Bigram perform similarly. SVM with Unigram TF-IDF Vectorization thus proves to be the better option for sentiment analysis on this dataset due to its lower error rate and higher accuracy.

Data:	Model Name	Incorrect Predictions
Authentic	MNB: Unigram TF Vector	7
Authentic	MNB: Unigram & Bigram	9
Authentic	MNB: Unigram TF-IDF	6
Authentic	SVM: Unigram TF	16
Authentic	SVM: Unigram & Bigram	19
Authentic	SVM: Unigram TF-IDF	19

The Multinomial Naive Bayes (MNB) model with Unigram Term Frequency Vectorization (MNB: Unigram TF Vector) exhibits the best performance with the fewest incorrect predictions among the authenticity classification models. With comparatively few errors, the MNB model with Unigram TF-IDF Vectorization also performs well. Support Vector Machine (SVM) models, especially those with Unigram & Bigram Term Frequency Vectorization and Unigram TF-IDF Vectorization, show higher percentages of incorrect predictions, indicating subpar performance for this particular task. The MNB: Unigram TF Vector model thus seems to be the most effective among the available options for authenticity classification.

e. A comparison of the difficulty in sentiment classification vs. fake review classification.

There are many levels of complexity for sentiment classification and fake review classification in text classification problems. Differentiating between good and negative sentiments in reviews is known as sentiment classification, and it is typically easier. The models created for this project's sentiment analysis performed well, with accuracy levels ranging from 75% to 96.4%. terms like "amazing" and "best" are frequently linked to positive sentiment, whereas terms like "bland" and "worst" are frequently linked to negative sentiment, making it possible to spot these patterns in text.

The process of classifying bogus reviews, which seeks to distinguish between real and phony reviews, proved to be more difficult. Other models struggled, with error rates ranging from 49% to 68%, and only the top-performing model managed to attain an accuracy of 79%. Since fake reviews frequently resemble real ones, it can be challenging to spot even little variations. Terms like "definitely," "recommend," and "kind," which may not always be unambiguous markers of genuineness, are frequently included in the differentiating characteristics for authenticity.

Because text contains clear markers of sentiment, sentiment classification is often thought to be less difficult than fake review classification, which necessitates a deeper examination and frequently involves more uncertainty.